Published in final edited form as:

J Neurosci Methods. 2019 March 15; 316: 3–11. doi:10.1016/j.jneumeth.2018.08.014.

A sleep spindle detection algorithm that emulates human expert spindle scoring

Karine Lacourse¹, Jacques Delfrate¹, Julien Beaudry¹, Paul Peppard², and Simon C. Warby^{1,3}

¹Center for Advanced Research in Sleep Medicine, Centre de Recherche de l'Hôpital du Sacré-Cœur de Montréal, Montreal, QC, Canada

²Population Health Sciences, University of Wisconsin School of Medicine and Public Health, Madison

³Département de Psychiatrie, Université de Montréal, Montreal, QC, Canada

Abstract

Background—Sleep spindles are a marker of stage 2 NREM sleep that are linked to learning & memory and are altered by many neurological diseases. Although visual inspection of the EEG is considered the gold standard for spindle detection, it is time-consuming, costly and can introduce inter/ra-scorer bias.

New Method—Our goal was to develop a simple and efficient sleep-spindle detector (algorithm #7, or 'A7') that emulates human scoring. 'A7' runs on a single EEG channel and relies on four parameters: the absolute sigma power, relative sigma power, and correlation/covariance of the sigma bandpassed signal to the original EEG signal. To test the performance of the detector, we compared it against a gold standard spindle dataset derived from the consensus of a group of human experts.

Results—The by-event performance of the 'A7' spindle detector was 74% precision, 68% recall (sensitivity), and an F1-score of 0.70. This performance was equivalent to an individual human expert (average F1-score = 0.67).

Comparison with Existing Method(s)—The F1-score of 'A7' was 0.17 points higher than other spindle detectors tested. Existing detectors have a tendency to find large numbers of false positives compared to human scorers. On a by-subject basis, the spindle density estimates produced by A7 were well correlated with human experts ($r^2 = 0.82$) compared to the existing detectors (average $r^2 = 0.27$).

Corresponding author: simon.c.warby@umontreal.ca, Centre d'études avancées en médecine du sommeil (CEAMS), 5400, boul. Gouin Ouest, Montréal (Québec) H4J 1C5.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Appendix A. Software Code

Code for the A7 spindle detector can be found online at <https://github.com/swarby>.

Conclusions—The ‘A7’ detector is a sensitive and precise tool designed to emulate human spindle scoring by minimizing the number of ‘hidden spindles’ detected. We provide an open-source implementation of this detector for further use and testing.

Keywords

Sleep; electroencephalography (EEG); polysomnography (PSG); sleep spindle; detector; sigma

1. Introduction

A sleep spindle is an electroencephalography (EEG) pattern that results from specific variations in membrane potentials in the thalamocortical network of the brain (De Gennaro and Ferrara, 2003). They are defined as “a train of distinct waves with frequency 11–16 Hz (most commonly 12–14 Hz) with duration 0.5 seconds, usually maximal in amplitude using central derivations”. (Iber and American Academy of Sleep Medicine, 2007) Spindles are a hallmark of stage 2 sleep as they define the transition from stage N1 (non-rapid eye movement, NREM1) to stage N2 (NREM2).

Spindle density and characteristics such as mean oscillation frequency, amplitude and duration appear to be trait-like, because they are stable over time (inter-night stability) for the same subject but vary considerably between subjects (Silverstein and Levy, 1976; Tan et al., 2001; Werth et al., 1997).

Although the exact function of the sleep spindle remains unclear, sleep spindle activity has clearly been associated with cognitive function such as intellectual ability, learning and memory consolidation (Diekelmann and Born, 2010; Walker, 2009; Fogel and Smith, 2011). In addition, alterations in the number of sleep spindles and their characteristics is linked to several sleep and mental disorders such as schizophrenia (Ferrarelli et al., 2007; Wamsley et al., 2012), neurodegenerative diseases (Christensen et al., 2014; Petit et al., 2004) and autism (Godbout et al., 1998; Limoges et al., 2005; Tessier et al., 2015).

Sleep spindles have typically been identified by visual inspection of the EEG signal but can also be accomplished by automated spindle detectors (De Gennaro and Ferrara, 2003). For human clinical polysomnography, the visual scoring of sleep spindles by human experts is generally considered as the gold standard, but it is time-consuming, costly and can introduce inter and intra-scorer bias (Danker-Hopfe et al., 2004; Silber et al., 2007; Wendt et al., 2015).

Automated spindle detectors resolve the subjectivity and efficiency problem of the human expert scoring but there are no standard guidelines for spindle detector properties. Spindle detection algorithms only have moderate agreement with each other, or with the human eye (Warby et al., 2014). Automated detections are highly influenced by the algorithm and settings chosen by researchers (Huupponen et al., 2007; Weiner and Dang-Vu, 2016). Even if many different algorithms are available (Warby et al., 2014; O’Reilly et al., 2016), none have been evaluated, in their original publication, against a group consensus of at least three human experts, in order to generalize the results (Warby et al., 2014; Wendt et al., 2015). Using a large sleep-spindle dataset created from a group consensus of experts (each epoch is

scored in average by 5 experts), we previously showed that six automated detectors only have moderate agreement with human scorers when estimating spindle density in middle-to-old-age subjects (Warby et al., 2014). The performance of the automated detectors was less than individual experts, or the group consensus of non-experts. In addition, spindle activity was better estimated with the relative sigma power than the automated detectors (Warby et al., 2014).

The purpose of this study is to provide a simple and efficient spindle detector that emulates human expert sleep spindle scoring. The current spindle detector design is unique because it uses a correlation filter between the EEG signal filtered in the sigma band and the raw EEG signal itself. The proposed design is therefore biased to detect spindles that are visible on the raw EEG signal by requiring a high correlation between raw EEG signal and the filtered sigma burst (the pattern that represents a spindle). We evaluate the performance of the A7 detector (algorithm #7) against the consensus of a group of human experts, as well as four other automated detectors.

2. Methods

2.1 EEG recording

The current study used the EEG recordings from a previous crowd-sourcing project of sleep spindles (Warby et al., 2014) that originated from the Wisconsin Sleep Cohort (WSC_{1950}) (Peppard et al., 2013). The WSC_{1950} originally had 2116 EEG recordings from 979 subjects of 57 ± 8 years old (47% female), but 166 recordings were excluded because of poor EEG quality or benzodiazepine drug use. The subsample cohort (named here WSC_{110sub}) used in the crowd-sourcing project includes 110 healthy subjects of 57 ± 8 years old (53% female). As previously (Warby et al., 2014), we used a random subset (WSC_{110sub}) of artifact-free stage N2 sleep from the EEG of these subjects: two blocks of 115 seconds for 100 subjects and 20 blocks of 115 seconds for the remaining 10 subjects. Channel C3-M2 was used to perform the spindle detection since the amplitude of the spindle is maximal at the central deviations (Iber and American Academy of Sleep Medicine, 2007). The EEG signal was band passed filtered 0.3–30 Hz (EEG_{bf} , ie the ‘raw EEG’ signal) (Supplementary Methods 2.1) according to standard practice for clinical polysomnography (Iber and American Academy of Sleep Medicine, 2007). In total, the WSC_{110sub} has around 13 hours of artifact-free N2 sleep.

2.2 Gold Standard (GS) definition

From WSC_{110sub} epochs of 25 sec were scored for sleep spindles by a total of 24 experts (qualified and experienced sleep technologists) following AASM guidelines and summary instructions from the crowdsourcing project (Warby et al., 2014, appendix). Each 25 sec epoch was reviewed by an average of five human experts. The group consensus is established and used as the gold standard (GS), as previously defined (Warby et al., 2014). Expert scored each spindle with a confidence score to indicate how confident they were in the spindle identification. The confidence scores were weighted as follows (one per spindle scored):

- ‘Definitely’ with a weight of 1

- ‘Probably’ with a weight of 0.75
- ‘Maybe’ with a weight of 0.5
- No spindle detection is weighted as 0

In order to produce a high-quality *GS* dataset, the spindle confidence scores (between 0 and 1) are averaged across all experts who viewed a specific epoch, resulting in a group consensus value (average of the confidence scores) for each sample in the epoch (i.e. the sample domain). Where 0 means all the experts who have seen the epochs agreed that there is no spindle, and 1 means all experts agreed that there is definitely a spindle. Values between 0 and 1 mean a partial agreement across experts. A spindle is found by the group when their consensus value exceeds the group consensus threshold (*GCT*). A *GCT* of 0.25 was chosen previously (Warby et al., 2014) because this value maximizes the mean individual expert performance and minimizes the standard deviation of the mean individual expert performance. The evaluation of the performance of each human expert was always against a group consensus in which the current expert did not contribute to the spindle scoring; i.e. a group consensus with the leave-one-out correction.

In order to standardize the spindles in the *GS*, short events (<0.3 sec) too close to each other (<0.1 sec) were merged together, and events were required to be longer than a minimum (0.3 sec) and shorter than a maximum duration (2.5 sec). This management of the spindle length was applied on the group consensus itself and not on each expert scoring. Despite the recommendation of a minimum event duration of 0.5 seconds (Iber and American Academy of Sleep Medicine, 2007), a 0.3 second minimum was chosen because human experts can reliably detect spindle as short as 0.3 seconds. These shorter spindles have the same characteristics (oscillation frequency, amplitude, symmetry) as longer spindles (Warby et al., 2014).

2.3 A7 definition

The proposed spindle detector is labelled A7 because 6 detectors (A1-A6) were previously evaluated (Warby et al., 2014). Based on the sleep spindle definition and standard practice used by human experts for sleep staging (Iber and American Academy of Sleep Medicine, 2007), the automated spindle detector should identify spindle events that are a distinct train of sigma (11–16 Hz) waves in the broadband signal (*EEGbf*; 0.3–30Hz).

Four parameters are extracted from the EEG signal to detect spindles. The broadband filtered EEG signals (0.3–30Hz; *EEGbf*) and the sigma band passed filtered (11–16 Hz; *EEGσ*) were used (Supplementary Methods 2.1). Both covariance and correlation were used to determine the relationship and measure the dependency between *EEGbf* and *EEGσ*. The four parameters are:

1. **Absolute sigma power** (*A7absSigPow*) to identify train of sigma waves (increase of power in the sigma band i.e. increase of energy of the signal *EEGσ*)
2. **Relative sigma power** (*A7relSigPow*) to ensure the increase of power is specific to the sigma band in the filtered signal *EEGbf*.

3. **Sigma covariance** (*A7sigmaCov*) to identify a high covariance between *EEGσ* and *EEGbf*. A high sigma covariance will indicate that *EEGσ* and *EEGbf* vary together.
4. **Sigma correlation** (*A7sigmaCorr*) to identify a high correlation between *EEGσ* and *EEGbf*. A high sigma correlation will indicate that the changes in *EEGσ* result in the change in *EEGbf*.

The four parameters described briefly above are computed on a 0.3 sec window length (*WinL*) each 0.1 sec for the whole EEG recording. (i.e. A sliding window (*WinL*) of 0.3 sec with a step window (*WinS*) of 0.1 sec is used). The 0.3 sec length is chosen to allow for detection of spindles as short as 0.3 sec length; the sliding window provides precision to 0.1 of a second for the event duration. A7's thresholds were computed from an artefact-free N2 sleep stage baseline.

The absolute sigma power is computed as the average of the squared value of each data sample of the *EEGσ* in the *WinL*. The absolute power is also log 10 transformed to normalize the distribution. The units are log 10(μV²). The *A7absSigPow* parameter is defined in the Equation 1.

$$A7absSigPow = \log_{10} \sum_{i=1}^N \frac{EEG\sigma_i^2}{N}$$

Equation 1 : The definition of A7 absolute sigma power parameter (*A7absSigPow*) where N is the number of samples in the window length (*WinL*), *i* is the index sample of the EEG signal band passed filtered in the sigma band (11–16 Hz) (*EEGσ*). The power is log 10 transformed.

The relative sigma power parameter is computed through a power spectral analysis (*PSA* - Supplementary Methods 2.2). *A7relSigPow* is defined as the ratio of the power in the sigma band over the broadband excluding the delta band (0.3–4.5 Hz). The power ratio is then log 10 transformed (*logPowRat*) and converted to a z-score (*A7relSigPow*) relative to the *logPowRat* distribution of 30-sec baseline of *EEGbf* centered on the current PSA window analyzed. The normalization with the z-score used a standard deviation computed using only the 10th to 90th percentile of the distribution of *logPowRat* to ignore extreme cases. The baseline includes only artifact-free data and may be restricted to specific sleep stages. The *A7relSigPow* is defined in the Equation 2; its unit is the z-score value.

$$A7relSigPow = zscore \left(\log_{10} \left(\frac{PSA_{11-16Hz}}{PSA_{4.5-30Hz}} \right) \right)$$

Equation 2: The definition of A7 relative sigma power parameter (*A7relSigPow*). The *PSA*_{11–16} Hz is the power in the sigma band and *PSA*_{4.5–30} Hz is the power in the broadband signal excluding delta band. The power ratio is log 10 transformed and converted to a z-

score relative the 30 sec of power ratio (*logPowRat*) around the current PSA window analyzed.

The **A7sigmaCov** parameter is based on the covariance between the broadband signal (*EEGbf*) and the signal filtered in sigma (*EEGσ*). *EEGbf* and *EEGσ* have the same number of samples (N), which is defined by the window length (*WinL*). Each sample in *EEGbf* has a paired (matched) sample in *EEGσ*. The covariance between *EEGbf* and *EEGσ* is the average product of the paired samples (Equation 3). Note that the variables are centered (their mean μ is removed) and the polarity (the sign) of each sample of the paired samples must be the same to give a positive covariance. Therefore, a high covariance results when *EEGbf* and *EEGσ* are varying together with the same polarity. The covariance values can vary from $-\infty$ to ∞ .

$$\text{cov}(\text{EEGbf}, \text{EEG}\sigma) = \frac{1}{N} \sum_{i=1}^N \left(\text{EEGbf}_i - \mu_{\text{EEGbf}} \right) * \left(\text{EEG}\sigma_i - \mu_{\text{EEG}\sigma} \right)$$

Equation 3 : The definition of the covariance between a variable *EEGbf* and *EEGσ*. μ expresses the mean of its respective variable, i is the sample index of each variable and N the total number of samples included in the variable.

The **A7sigmaCov** parameter is the covariance between *EEGbf* and *EEGσ* log 10 transformed (*logCov*) and converted to a z-score relative to the *logCov* distribution of 30 sec baseline of *EEGbf* around the current window analyzed. The z-score normalization uses a standard deviation computed only on the 10th to 90th percentile of the distribution of *logCov* to ignore extreme values. The baseline includes data without EEG artifacts and may be limited to specific sleep stages when desired. The **A7sigmaCov** parameter is defined in Equation 4, its unit is the z-score value.

$$\text{A7sigmaCov} = \text{zscore} \left(\log_{10} \left(\text{cov} \left(\text{EEG}_{bf}, \text{EEG}\sigma \right) \right) \right)$$

Equation 4 : The definition of A7 sigma covariance parameter (*A7sigmaCov*) where *cov* is the covariance between the EEG signals filtered 0.5–30 Hz (*EEGbf*) and the EEG signal filtered in the sigma band (*EEGσ*).

The **A7sigmaCorr** parameter is the correlation between *EEGσ* and *EEGbf*. The correlation is defined as their covariance normalized by the amplitude of both variables. The normalization is simply a division by the product of the standard deviations (*sd*) of the two variables. The **A7sigmaCorr** parameter is defined in Equation 5. The values of *A7sigmaCorr* range from -1 to 1 , but mostly between 0 and 0.5 , which typically results in a normal distribution of values and is therefore not log transformed.

$$A7sigmaCor = \frac{cov(EEG_{bf}, EEG_{\sigma})}{sd_{EEG_{bf}} * sd_{EEG_{\sigma}}}$$

Equation 5 : The definition of A7 sigma correlation parameter (*A7sigmaCorr*) where *cov* is the covariance and *sd* is the standard deviation of *EEG_{bf}* and *EEG_σ*.

1. The four A7 parameters (*A7absSigPow*, *A7relSigPow*, *A7sigmaCov* and *A7SimaCor*) are computed on a 0.3 sec window length (*WinL*) each 0.1 sec (*WinS*) for the whole EEG recording.
2. An event is detected when the four parameters exceed their respective thresholds.
3. The length of the spindle is based on two A7 parameters: “*A7absSigPow* and *A7sigmaCov*”. I.e. Every consecutive 0.1 sec windows around the detected event are considered part of the spindle if the *A7absSigPow* and *A7sigmaCov* exceed their respective threshold. The time resolution of the detections is therefore 0.1 sec.
4. Events that are < 0.3 seconds or > 2.5 seconds are discarded.

2.4 Context classifier

A spindle context classifier is included in the design of the A7 detector. The purpose of the classifier is to label each detected spindle as occurring “IN” or “OUT” the expected spectral context for spindles. The idea is to identify 30 sec window lengths with a spectral profile similar to N2-N3 sleep stages where spindles should be found most typically (i.e. NREM sleep). The context classifier is especially useful when the sleep staging is not available or reliable, although it has not been specifically designed to replace the sleep staging. The classifier does not modify the detection procedure directly, but is used to classify spindles that the A7 algorithm detects. For testing the context classifier, we used the WSC₁₉₅₀ dataset (all sleep stages), and since the baseline used to compute the *A7relSigma* and the *A7sigmaCov* parameters was no longer limited to N2, their baseline was sleep staging independent. Spindles ‘OUT’ of context can be removed from the analysis, or selectively quantified, depending on the research question.

The context classification is based on the power ratio of the sum of the delta and theta over the beta band (*slowRatio* = 0.5–8 Hz / 16–32 Hz) through the PSA described in Supplementary Methods section 2.2. The procedure to label each spindle as “IN” or “OUT” the expected spectral context for spindles is described below:

1. The PSA is performed through a sliding window of 0.3 sec length (*WinL*) each 0.1 sec (*WinS*) for the whole recording in order to estimate the power of different frequency bands.
2. The average slow power (0.5–8 Hz; *slowAvgBs*), and the average fast power (16–32 Hz; *fastAvgBs*), are computed for a clean (without any artifacts) 30 sec window centered on each PSA window.

3. The average *slowRatio* ($slowAvgBsl / fastAvgBsl$) distribution is normalized with a log 10 transformation ($log_{10}SlowRatio$).
4. A detected event is classified “IN” the expected spectral context for spindles if it occurs in a 30 sec window with a $log_{10}SlowRatio >$ than the slow ratio threshold (*sRatThresh*).

2.5 Performance analysis

The performance of the spindle detectors is determined by comparison to the GS, a set of EEG recordings with spindles identified from the consensus of human experts scoring (described in the section 2.2) After thresholding the group consensus with a GCT of 0.25, the GS becomes a binary vector of samples indicating the location of the spindles in the time series.

A by-subject analysis (a single value per subject) is sufficient to estimate the spindle density per subject, however, characteristics of the spindles detected, such as duration, amplitude or oscillation frequency are relevant. In order to determine these characteristics, a by-event (spindle-by-spindle) or by-sample (sample point-by-sample point in the time series) analysis of the EEG time series is required. A by-event analysis considers the spindle itself as a measurement unit and is therefore the most useful for our study. The by-sample performance is however also available in Supplementary Section 3.4.

Spindles can be variable in length and therefore can lead to imperfect matching between the detected event and spindle event in the GS (Warby et al., 2014). To consider a spindle as correctly detected the algorithm detection (D) and the GS spindle event (E) must overlap (O_{ED}) above certain overlap threshold (OvT). The O_{ED} is defined in Equation 6 and is the intersection over the union between E and D .

$$O_{ED} = \frac{E \cap D}{E \cup D}$$

Equation 6 : The definition of the Overlap (O_{ED}) between detection (D) and an event (E) where $\cap$ is the intersection and $\cup$ the union between E and D . The $\cap$ is the part of E detected and $\cup$ the sum of the length of E and D .

For performance evaluation, we used an OvT of 0.2 to maximize the by-event performance of the experts and automated methods (Supplementary Figure 2). Multiple overlaps are not allowed; only one D can match an E . When multiple overlaps are possible, the best match is determined by the ED pair with the maximum O_{ED} defined in Equation 6.

The performance of each detector against the GS is described with the number of correct detections called (True Positive; TP), the incorrect detection called (False Positive; FP) and the events missed called (False Negative; FN). The performance is expressed in this study with these three following definitions:

- **Recall:** the proportion of spindles detected $\left(\frac{TP}{TP + FN} \right)$

- **Precision:** the proportion of detected events considered as real spindle $\left(\frac{TP}{TP + FP}\right)$
- **F1-score:** the harmonic mean of the recall and precision $\left(2 \frac{Precision * Recall}{Precision + Recall}\right)$

The recall (sensitivity), precision and F1-score are chosen to describe the performance because they are not based on the correct ‘non-detections’ called (True Negative; TN). The number of TN is high when the events, such as spindles, are rare in the dataset. With a very high number of TN, metrics such as specificity are not useful, as they will always be very high, even if the overall performance is very low.

2.6 Cross Validation

In order to avoid an A7 design that over-fits the GS dataset, we split the WSC_{110sub} into a training and validation dataset. The training dataset is used to establish the best set of A7 thresholds for its four parameters. The validation dataset is used to test the thresholds computed from the training dataset without modifying them. The WSC_{110sub} was divided using a 50/50 split by randomly selecting 200 blocks from the 400 blocks of 115 sec of artifact-free N2 available for the training dataset and the remaining 200 blocks were the validation dataset. Each dataset has around 6 h 30 mins of N2 sleep without artifacts.

The training and validation datasets were also bootstrapped independently in order to evaluate the stability of the performance across sets of blocks. One sequence of bootstrap is a random selection without replacement of 40 blocks out of the 200 available. Five sequences of bootstrap are applied to each dataset; each sequence therefore includes around 1 h 15 min of N2.

The thresholds of the A7 parameters (*A7absSigPow*, *A7relSigPow*, *A7sigmaCov* and *A7SigmaCorr*) are optimized to maximize the by-event F1-score (F1-score_{max}) with an *OvT*=0.5 on the training dataset. Note that a stricter overlap threshold (*OvT*=0.5) is used for the training phase than the validation and performance evaluation (*OvT*=0.2) in order to achieve a better estimation of the detected spindle duration.

To determine whether the characteristics of the spindle detections are similar to the characteristics of the GS events, we evaluated and compared three spindle characteristics (Table 1).

In addition to evaluating the performance of A7 against the GS, we compared four additional spindle detectors as a reference: A2 (Ferrarelli et al., 2007), A3 (Möller et al., 2002), A4 (Martin et al., 2013) and A5 (Wamsley et al. 2012), (Supplementary Table 1). These detectors were selected because only one EEG channel is needed to perform the spindle detection. The details of how these detectors function is found in the original publications, and summarized previously (Warby et al., 2014). The performance of the A2-A5 detectors and experts in this study are similar to those reported previously (Warby et al., 2014); some differences are discussed in the Supplementary Results section 3.

3. Results

3.1 A7 Parameters analysis

The A7 algorithm detects a spindle when the four A7 sigma parameters exceed their respective thresholds. These four thresholds are established with the training dataset from WSC_{110sub}.

Table 2 presents the set of thresholds that maximized the by-event F1-score with an overlap threshold (*OvT*) of 0.5. The procedure and range of values tested to obtain a set of thresholds are presented in Supplementary Methods section 2.5.

Figure 1 is an example of 4 spindle detections in a 25 sec epoch. The four A7 parameters *A7absSigPow*, *A7relSigPow*, *A7sigmaCov* and *A7sigmaCorr* synchronously exceed their respective thresholds during the detections. The windows where the four parameters first exceed their threshold is considered the origin of the detection. The detected event duration is extended for every consecutive 0.1 sec window (*WinS*) around the origin where both *A7absSigPow* and *A7sigmaCov* continue to exceed their threshold.

3.2 A7 Performance analysis

The F1-score for the automated detectors ranged from 0.27 to 0.70 (Table 3). The A7 detector had the highest F1-score 0.70 ± 0.03 , which was 0.17 points higher than the next automated detector and 0.03 points higher than the average individual human expert compared to the expert group.

Many of the individual experts, as well as some automated detectors, show a bias for precision, at the cost of recall (or vice versa), while other experts/detectors find a balance between the two, including the A7 detector (Figure 2; diagonal line).

The performance of all of the detectors is stable for an overlap threshold up to 0.4. However, for more stringent overlap thresholds (*OvT* > 0.4), by-event performance begins to decay in a similar fashion at higher thresholds for all detectors (Supplementary Section 3.3). Analyzing performance at the by-sample level also provides similar results to the by-event analysis (Supplementary Section 3.4).

The complete WSC_{110sub} (training + validation) is used to compute the spindle density because there was similar performance between the training and validation datasets (Supplementary Table 4). The GS (the human expert consensus) computed from the whole WSC_{110sub} includes 1987 spindles. The A7 spindle detector identified a total of 1938 events: 582 were incorrectly detected (FP), 631 spindles from the GS were missed (FN), and 1356 spindles were correctly detected (TP). The spindle density computed on the events detected by A7 is the most similar to the spindle density from the GS (expert consensus on WSC_{110sub}). The average density is 2.32 spm for the GS and 2.57 spm for the A7. Figure 3 shows the spindle density distribution computed on the events of the GS and events of each automated detector. Average spindle density for all detectors is summarized in the Supplementary Table 3.

The correlation of the spindle density between the GS events and the detected events is shown, for each detector tested, in Figure 4. The highest correlation is between GS and the A7 detector ($R^2=0.82$).

We evaluated the mean spindle characteristics (Table 1) of the GS events, the A7 detections, the GS events missed by the A7 detector ($GS-TP=FN$) and the spindle incorrectly identified (bad detections) by the A7 ($A7-TP=FP$) (Table 4). Overall, the mean characteristics of the A7 spindle detections are similar to the GS. The spindles missed (FN) are 0.07 sec shorter and 3.28 μV lower in amplitude than the GS. The bad detections (FP) are also on average shorter (0.15 sec) and of lower amplitude (1.45 μV).

3.3 Spindle Context Classifier

While the performance reported above does not use the context classifier, the results could be further refined by classifying each detected spindle event as being “IN” or “OUT” of the expected spectral context for spindles (see section 2.4 for more details). In our study, the PSA-based threshold *slRatThresh* to classify 30 sec baseline around each spindle is based on the $\log_{10}SlowRatio$ distribution of all the epochs (30 sec length) of the *WSC₁₉₅₀*. A *slRatThresh* of 0.9 distinguishes around half of the epochs scored awake from almost all the epochs scored N2. Figure 5 shows the probability histogram of the $\log_{10}SlowRatio$ of the epochs of the *WSC₁₉₅₀* for each sleep stage. The resulting contextual classifications (“IN”/“OUT”) can be used as another characteristic of the spindle and quantified (i.e. quantitating the number of spindles in an unusual context, for example in the case of disease or drug effects), or as a filtering criteria to remove spindle detections that are more likely to have occurred during wake.

To determine the usefulness of the context classifier, we evaluated the all night EEG recordings of *WSC₁₉₅₀* (by-EEG recording analysis). Using the context classifier to remove spindle events “OUT” of the expected spectral context selectively reduced the number of spindle detections that are found outside of NREM2/3 sleep (Table 5). Moreover, the correlation between number of spindles detected all night but “IN” context and the total number of spindles detected in N2 only was $R^2=0.93$ ie “IN” context spindles will be a good representation of N2 spindles in the absence of sleep staging (Supplementary Figure 3). Example of an N2 epoch “OUT” of context is in the supplementary results (Supplementary Figure 5).

4. Discussion

In this study we propose a new spindle detection algorithm that relies on four parameters extracted from the raw and sigma band-passed filtered EEG signal: *A7absSigPow*, *A7relSigPow*, *A7sigmaCov* and *A7sigmaCorr*.

Using this approach, the A7 detector was able to identify spindles in a manner similar to individual human experts in a cohort from middle- and older-aged subjects. Compared to a human expert-generated GS, the A7 detector produced the highest F1-score (0.70), the highest spindle density correlation of individual subjects ($r^2=0.82$), and average spindle characteristics that are closest to the GS of all the automated spindle detectors we tested. For

F1-score, the A7 detector performed comparably to individual human experts. The A7 detector, like the A5 and A4 detectors, was well-balanced between the competing metrics of recall and precision. In contrast, detectors A2 and A3 would be useful for applications that require maximizing precision or recall respectively, rather than the balance between the two.

4.1 Visual scoring versus automated detector

There are several cases where human scored spindles have produced different results than existing algorithms. Some results of studies regarding spindle activity differences between clinical and nonclinical groups are inconsistent (Gruber and Wise, 2016; Weiner and Dang-Vu, 2016). For instance, in autism patients, previous studies (Godbout et al., 1998; Limoges et al., 2005; Miano et al., 2004; Tessier et al., 2015), found a significant decrease in sleep spindle density using visual scoring, whereas Tani et al. (2004) and Sahroni et al. (2016) did not find any difference using an automatic detector. Decreases in spindle density have been also observed in narcolepsy with visual scoring (Delrosso et al., 2014) but not replicated with an automated detector (Christensen et al., 2017). In contrast, automated detectors have found differences in spindle density in insomnia symptoms (Dang-Vu et al., 2015), but manual scoring of spindles have found no difference in spindle density in insomnia patients (Bastien et al., 2009). The spindle activity over the course of a night is also described differently depending on the method analysis used. While visual scoring has unequivocally shown an increase density over the course of the night, there is no consensus with automated detectors (De Gennaro and Ferrara, 2003).

4.2 A7 spindle detector contribution

While it is clear that automated methods provide numerous advantages (scalability, reliability, speed) over visually scored spindles, the differences in previous research findings outlined above suggest that there is a need for an automated method that emulates visual scoring. Automated spindle detectors have a tendency to have very high false-positive rates relative to visual scoring. These “hidden spindles” that cannot be seen in the raw EEG (*EEGbf*) are found in all stages of sleep and wake, and their inclusion as true spindles is therefore problematic unless we change the definition of sleep spindles so that they are no longer a NREM2/3, or even sleep-specific phenomena. The problem is also compounded by the fact that there is a lack of consensus among automated detectors; they all find different subsets of “hidden spindles” (Warby et al., 2014). Human scoring of spindles (particularly for sleep staging) is generally performed on the *EEGbf* signal, with an emphasis on spindles that can be distinguished visually from their surroundings. Of course, the value of human-scored spindles as a gold standard can be disputed. However, at the moment, this is the best gold standard we have for evaluating the performance of spindle detectors, and was used to develop the A7 detector.

4.3 Spectral context for spindles

We also propose the use of a context classifier to help identify or eliminate events that meet the criteria for a spindle, but are found in an EEG context (i.e. 30 seconds of EEG around the spindle event) that is not typical for spindles. By considering only the spindles classified “IN” spectral context for the A7 algorithm, the percentage of detections in the sleep stages where the events should not be typically found is decreased. We observed a reduction of

83% in awake and unscored epochs whereas the reduction was only of 8% in N2 and 0.4% in N3. Moreover, number of spindles detected “IN” context, without using any input sleep staging, was highly correlated to the number of spindles in stage N2. However, while “OUT” of context spindles could be just removed from the analysis, they also may be a useful phenotype in themselves, and “OUT” of context spindles may be enriched by certain drugs or diseases. This remains to be determined. The A7 context classifier relies on the *sIRatThresh*. Since the slow ratio threshold (*sIRatThresh*=0.9) is based on a middle-age cohort of almost 2000 recordings it is probably a decent representation of a middle-old age cohort. However, *sIRatThresh* should be evaluated carefully in a cohort with younger subjects or patients with disease.

4.4 Signal-to-noise ratio of spindles

The A7 algorithm emulates human scoring by requiring that the signal found in the sigma band-passed data (*EEG σ* ; 11–16Hz) retains a (visual) relationship with the original *EEGbf* (0.3–30Hz) data series using the *A7sigmaCov* and *A7sigmaCorr* parameters. In other words, the sigma power used to detect spindles must be somewhat visible in the original EEG signal. If spindles are completely obscured in noise (where ‘noise’ is defined as any signal other than spindles) they will not be detected by the A7 algorithm. In this way, the A7 detector has a bias towards spindles that have a good signal-to-noise ratio. The implications of this are unclear, as the value of hidden spindles, that can be found by algorithms but not seen in the original EEG (and would be missed by humans in a typical scoring scenario) is unknown. However, we would put forward the hypothesis that there is a relationship between the signal-to-noise ratio in the EEG, and functional effectiveness of spindles. Spindles perform some functions (i.e. learning and memory consolidation) that are aging related, and spindles with a poor signal-to-noise ratio may perform this function less well. If this is the case, then detecting spindles with a good signal-to-noise ratio (as human scorers do) has value. As spindle density and other characteristics specifically decrease with age, the signal-to-noise ratio decreases. This remains to be determined experimentally. While we used cross validation to establish reliable thresholds in an older subject population, this may need to be re-evaluated in younger subjects.

Spindles may have different characteristics depending of the locations on the scalp. However, A7 was developed at C3 only. Thus, the performance of A7 algorithm should be assessed with other input channels. Even though spindle characteristics and density change between subjects and are age dependent, the data distribution is often around a single mean value. Therefore by adapting the thresholds to the age of the population study, the methodology implemented in this paper should remain still valid. For the detection of ‘fast’ and ‘slow’ spindles, we suggest that this can easily done after the detection, by classifying spindles as ‘fast’ and ‘slow’ based on their oscillation frequency, if desired.

Conclusion

The A7 detector was developed to emulate human sleep spindle scoring and was tested against a human scored gold standard generated by numerous human experts. While there is a clearly a need for more refined definitions and standardization of sleep spindle detection,

we propose that emulating the way the human eye detects sleep spindles may be useful for some (but not all) research questions. There are numerous methodological approaches to sleep spindle detection, and each may have specific strengths and weaknesses depending on the application.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to acknowledge the contributions of the Wisconsin Sleep Cohort and thank the subjects for their participation. Funding for this work was provided by the Chaire Pfizer, Bristol-Myers Squibb, SmithKline Beecham, Eli Lilly en psychopharmacologie de l'Université de Montréal, and the Canadian Institutes of Health Research (CIHR - OOGP 313177). The Wisconsin Sleep Cohort Study was supported by the National Heart, Lung, and Blood Institute (R01HL62252) and the National Center for Research Resources (1UL1RR025011) at the US NIH.

Abbreviation table

AASM

American Academy of Sleep Medicine

A7absSigPow

Absolute sigma power parameter for A7 detector

A7relSigPow

Relative Sigma power parameter for A7 detector

A7sigmaCorr

Sigma correlation parameter for A7 detector

A7sigmaCov

Sigma covariance parameter for A7 detector

By-event analysis/performance

spindle-by-spindle analysis/performance

By-sample analysis/performance

sample point-by-sample point analysis/performance

Detectors tested – algorithm *A*

A2: Ferrarelli et al. 2007

A3: Mölle et al. 2002

A4: Martin et al. 2013

A5: Wamsley et al. 2012

A7: currently proposed detector

EEG

Electroencephalography

EEG_{bf}

EEG broadband signal (filtered 0.3–30 Hz)

EEG σ

EEG sigma signal (band-passed filtered 11–16 Hz)

fastAvgBsl

Average fast power (16–32 Hz)

FFT

Fast Fourier transform

FN

False negative

FP

False positive

GCT

Group consensus threshold

GS

Gold Standard

IExp

Individual human experts

IntegSpectPow

Integration of spectral power

NF

Nyquist frequency

NG

Noise gain

NREM

Non rapid eye movement

O_{ED}

Overlap between the GS spindle event E and the algorithm detection D

O_{vT}

Overlap threshold

PSA

Power spectral analysis

PSG

Polysomnography

slowAvgBsl

Average slow power (0.5–8 Hz)

slowRatio

slowAvgBsl/fastAvgBsl

slRatThresh

Slow ratio threshold

SOS

Second Order Section

spm

Spindles per minute

TN

True negative

TP

True positive

WinL

Window length to compute A7 thresholds

WinS

Window step between two WinL

WSC₁₉₅₀

Wisconsin Sleep Cohort

WSC_{110sub}

Wisconsin Sleep Cohort subset (110 subjects)

References

- Bastien CH, St-Jean G, Turcotte I, Morin CM, Lavallée M, Carrier J, 2009 Sleep spindles in chronic psychophysiological insomnia. *J. Psychosom. Res* 66, 59–65. 10.1016/j.jpsychores.2008.05.013 [PubMed: 19073295]
- Christensen JAE, Kempfner J, Zoetmulder M, Leonthin HL, Arvastson L, Christensen SR, Sorensen HBD, Jennum P, 2014 Decreased sleep spindle density in patients with idiopathic REM sleep behavior disorder and patients with Parkinson's disease. *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol* 125, 512–519. 10.1016/j.clinph.2013.08.013
- Christensen JAE, Nikolic M, Hvidtfelt M, Kornum BR, Jennum P, 2017 Sleep spindle density in narcolepsy. *Sleep Med.* 34, 40–49. 10.1016/j.sleep.2017.02.022. [PubMed: 28522097]
- Dang-Vu TT, Salimi A, Boucetta S, Wenzel K, O'Byrne J, Brandewinder M, Berthomier C, Gouin J-P, 2015 Sleep spindles predict stress-related increases in sleep disturbances. *Front. Hum. Neurosci* 9, 68 10.3389/fnhum.2015.00068. [PubMed: 25713529]

- Danker-Hopfe H, Kunz D, Gruber G, Klösch G, Lorenzo JL, Himanen SL, Kemp B, Penzel T, Röschke J, Dorn H, Schlögl A, Trenker E, Dorffner G, 2004 Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J. Sleep Res* 13, 63–69. [PubMed: 14996037]
- De Gennaro L, Ferrara M, 2003 Sleep spindles: an overview. *Sleep Med. Rev* 7, 423–440. [PubMed: 14573378]
- Delrosso LM, Chesson AL, Hoque R, 2014 Manual characterization of sleep spindle index in patients with narcolepsy and idiopathic hypersomnia. *Sleep Disord.* 2014, 271802 10.1155/2014/271802
- Diekelmann S, Born J, 2010 The memory function of sleep. *Nat. Rev. Neurosci* 11, 114–126. 10.1038/nrn2762 [PubMed: 20046194]
- Ferrarelli F, Huber R, Peterson MJ, Massimini M, Murphy M, Riedner BA, Watson A, Bria P, Tononi G, 2007 Reduced sleep spindle activity in schizophrenia patients. *Am. J. Psychiatry* 164, 483–492. 10.1176/appi.ajp.164.3.483 [PubMed: 17329474]
- Fogel SM, Smith CT, 2011 The function of the sleep spindle: a physiological index of intelligence and a mechanism for sleep-dependent memory consolidation. *Neurosci. Biobehav. Rev* 35, 1154–1165. 10.1016/j.neubiorev.2010.12.003 [PubMed: 21167865]
- Godbout R, Bergeron C, Stip E, Motttron L, 1998 A Laboratory Study of Sleep and Dreaming in a Case of Asperger's Syndrome. *Dreaming* 8, 75–88. 10.1023/B:DREM.0000005898.95212.58
- Gruber R, Wise MS, 2016 Sleep Spindle Characteristics in Children with Neurodevelopmental Disorders and Their Relation to Cognition. *Neural Plast.* 2016. 10.1155/2016/4724792
- Huupponen E, Gómez-Herrero G, Saastamoinen A, Värri A, Hasan J, Himanen S-L, 2007 Development and comparison of four sleep spindle detection methods. *Artif. Intell. Med* 40, 157–170. 10.1016/j.artmed.2007.04.003 [PubMed: 17555950]
- Iber C, American Academy of Sleep Medicine, 2007 The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. American Academy of Sleep Medicine.
- Limoges E, Motttron L, Bolduc C, Berthiaume C, Godbout R, 2005 Atypical sleep architecture and the autism phenotype. *Brain J. Neurol* 128, 1049–1061. 10.1093/brain/awh425
- Martin N, Lafortune M, Godbout J, Barakat M, Robillard R, Poirier G, Bastien C, Carrier J, 2013 Topography of age-related changes in sleep spindles. *Neurobiol. Aging* 34, 468–476. 10.1016/j.neurobiolaging.2012.05.020 [PubMed: 22809452]
- Miano S, Bruni O, Leuzzi V, Elia M, Verrillo E, Ferri R, 2004 Sleep polygraphy in Angelman syndrome. *Clin. Neurophysiol* 115, 938–945. 10.1016/j.clinph.2003.11.004 [PubMed: 15003776]
- Mölle M, Marshall L, Gais S, Born J, 2002 Grouping of Spindle Activity during Slow Oscillations in Human Non-Rapid Eye Movement Sleep. *J. Neurosci* 22, 10941–10947. [PubMed: 12486189]
- O'Reilly C, Warby SC, Nielsen T, 2016 Editorial: Sleep Spindles: Breaking the Methodological Wall. *Front. Hum. Neurosci* 10, 672 10.3389/fnhum.2016.00672 [PubMed: 28149273]
- Peppard PE, Young T, Barnet JH, Palta M, Hagen EW, Hla KM, 2013 Increased prevalence of sleep-disordered breathing in adults. *Am. J. Epidemiol* 177, 1006–1014. 10.1093/aje/kws342 [PubMed: 23589584]
- Petit D, Gagnon J-F, Fantini ML, Ferini-Strambi L, Montplaisir J, 2004 Sleep and quantitative EEG in neurodegenerative disorders. *J. Psychosom. Res* 56, 487–496. 10.1016/j.jpsychores.2004.02.001 [PubMed: 15172204]
- Sahroni A, Igasaki T, Murayama N, Yudiyanta, 2016 Sleep Spindle Analysis on Typically Developing and Autistic Children During Sedation. *Neurosci. Biomed. Eng* 4, 202–208.
- Silber MH, Ancoli-Israel S, Bonnet MH, Chokroverty S, Grigg-Damberger MM, Hirshkowitz M, Kapen S, Keenan SA, Kryger MH, Penzel T, Pressman MR, Iber C, 2007 The visual scoring of sleep in adults. *J. Clin. Sleep Med. JCSM Off. Publ. Am. Acad. Sleep Med* 3, 121–131.
- Silverstein LD, Levy CM, 1976 The stability of the sigma sleep spindle. *Electroencephalogr. Clin. Neurophysiol* 40, 666–670. [PubMed: 57053]
- Tan X, Campbell IG, Feinberg I, 2001 Internight reliability and benchmark values for computer analyses of non-rapid eye movement (NREM) and REM EEG in normal young adult and elderly subjects. *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol* 112, 1540–1552.

- Tani P, Lindberg N, Nieminen-von Wendt T, von Wendt L, Virkkala J, Appelberg B, Porkka-Heiskanen T, 2004 Sleep in young adults with Asperger syndrome. *Neuropsychobiology* 50, 147–152. 10.1159/000079106 [PubMed: 15292669]
- Tessier S, Lambert A, Chicoine M, Scherzer P, Soulières I, Godbout R, 2015 Intelligence measures and stage 2 sleep in typically-developing and autistic children. *Int. J. Psychophysiol. Off. J. Int. Organ. Psychophysiol* 97, 58–65. 10.1016/j.ijpsycho.2015.05.003
- Walker MP, 2009 The role of sleep in cognition and emotion. *Ann. N. Y. Acad. Sci* 1156, 168–197. 10.1111/j.1749-6632.2009.04416.x [PubMed: 19338508]
- Wamsley EJ, Tucker MA, Shinn AK, Ono KE, McKinley SK, Ely AV, Goff DC, Stickgold R, Manoach DS, 2012 Reduced sleep spindles and spindle coherence in schizophrenia: mechanisms of impaired memory consolidation? *Biol. Psychiatry* 71, 154–161. 10.1016/j.biopsych.2011.08.008 [PubMed: 21967958]
- Warby SC, Wendt SL, Welinder P, Munk EGS, Carrillo O, Sorensen HBD, Jennum P, Peppard PE, Perona P, Mignot E, 2014 Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nat. Methods* 11, 385–392. 10.1038/nmeth.2855 [PubMed: 24562424]
- Weiner OM, Dang-Vu TT, 2016 Spindle Oscillations in Sleep Disorders: A Systematic Review. *Neural Plast.* 2016, 7328725. 10.1155/2016/7328725
- Wendt SL, Welinder P, Sorensen HBD, Peppard PE, Jennum P, Perona P, Mignot E, Warby SC, 2015 Inter-expert and intra-expert reliability in sleep spindle scoring. *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol* 126, 1548–1556. 10.1016/j.clinph.2014.10.158
- Werth E, Achermann P, Dijk DJ, Borbély AA, 1997 Spindle frequency activity in the sleep EEG: individual differences and topographic distribution. *Electroencephalogr. Clin. Neurophysiol* 103, 535–542. [PubMed: 9402884]

Highlights

This section consists of a short collection of bullet points that convey the core findings of the article and should be submitted in a separate editable file in the online submission system.

- The A7 ('algorithm #7') detector uses 4 parameters to identify spindles that are visible in the original EEG (ie not 'hidden' spindles).
- A7 emulates human spindles scoring by maximizing performance against a human gold standard.
- The 'context classifier' filters out spindles that are not in a typical NREM spectral context.
- Identifying visible spindles may be useful because they have a high signal-to-noise ratio.

Figure 1 :

The four A7 parameters: “ $A7absSigPow$, $A7relSigPow$, $A7sigmaCov$ and $A7sigmaCorr$ ” extracted from 25 sec of EEG signal band-pass filtered 0.3–30 Hz ($EEGbf$). (A) $EEGbf$ with three detected spindles marked in blue; (B) $EEGbf$ band-pass filtered in the sigma band 11–16 Hz (black); (C) $A7absSigPow$ in blue and $A7relSigPow$ in red with their respective thresholds in dash lines; (D) $A7sigmaCov$ in blue and $A7sigmaCorr$ in red with their respective thresholds in dash lines.

Figure 2: Precision-Recall plots of by-event performance of detectors and experts in the WSC_{110sub} validation dataset. Only the average through sequences of bootstrap is shown. For the human experts, the size of the circle indicates the amount of data they scored.

Figure 3 :

The spindle density distribution across subjects computed on the events of the gold standard (GS) and each detector (A2-A7) in WSC_{110sub} . The black cross indicates the median of the distribution (horizontal) and the interquartile range (vertical). The red dotted line represents the median spindle density per minute (spm) of the GS as a reference. Note that A7 has the closest median spm to the GS, but also has an overall distribution of spm most similar to the GS.

Figure 4 :

The correlation of the spindle density between the gold standard spindles per minute (GS spm) (x-axis) and the detected events (y-axis) for each detector (A2-A7). Each dot in the plot represents one subject; darker points indicate multiple subjects at the same position. The black bold line is the linear regression across subjects. The WSC_{110sub} is used. The linear regression p-value was significant (< 0.05) for each detector.

Figure 5 :

The probability histogram of the slow ratio log 10 transformed ($\log_{10}SlowRatio$) computed as the power of the delta and theta (0.5–8 Hz) over the beta (16–32 Hz) band. The dash black line shows the selected threshold ($slRatThresh$) of 0.9 for the spindle context classifier. WSC_{1950} is used.

Table 1:

Spindle characteristics evaluated to compare the A7 detections to the spindles of the gold standard (GS).

Label	Description	Units
durlnSec	Duration from the onset to the end of the spindle	sec
oscFreq	Oscillation frequency as the average 1/(peak-to-peak interval in sec) in the 11–16 Hz band-pass filtered signal ¹	Hz
maxPeak2PeakAmp	Maximum peak-to-peak amplitude of spindle in the 11–16 Hz band-pass filtered signal	μV

Table 2:

The threshold of each A7 sigma parameter that maximize the by-event F1-score. Thresholds were obtained using the training dataset of WSC_{110sub} and an overlap threshold (OvT) of 0.5.

A7 parameter	<i>A7absSigPow</i>	<i>A7relSigPow</i>	<i>A7sigmaCov</i>	<i>A7sigmaCor</i>
units	log 10(μV^2)	z-score	z-score	from -1 to 1
value	1.25	1.6	1.3	0.69

Table 3:

By-event performance of automated detectors (A2-A7) and individual human experts (IExp). The performance of the detectors is presented as the mean and the standard deviation across the five sequences of bootstrap. The performance of the individual experts is presented as the mean and standard deviation across the group of experts (24 experts, no bootstrap). The overlap threshold (OvT) used is 0.2. Performance was evaluated using the validation dataset of WSC_{110sub}.

Detector	Recall	Precision	F1-score
A2	0.17 ± 0.03	0.71 ± 0.05	0.27 ± 0.05
A3	0.77 ± 0.05	0.35 ± 0.02	0.49 ± 0.02
A4	0.57 ± 0.06	0.49 ± 0.04	0.53 ± 0.05
A5	0.56 ± 0.07	0.51 ± 0.04	0.53 ± 0.05
A7	0.68 ± 0.06	0.74 ± 0.04	0.70 ± 0.03
IExp	0.68 ± 0.19	0.71 ± 0.19	0.67 ± 0.13

Table 4:

The mean spindle characteristics of the gold standard (GS; consensus of the human experts), A7 detections, false negative (FN) and false positive (FP) in the WSC_{110sub} dataset.

Spindle characteristics	GS	A7	FN=GS-TP	FP=A7-TP
durlnSec (sec)	0.75	0.75	0.68	0.60
oscFreq (Hz)	13.54	13.50	13.40	13.33
maxPeak2PeakAmp (μV)	25.49	26.83	18.96	22.68

Table 5:

Mean ($\pm$ standard deviation) number of spindles per sleep stage, expressed as a percentage of all sleep stages. A7* detections are classified “IN” context only. The A7 detector with the context classifier (A7*) finds fewer spindles in wake (where spindles would not normally be expected to be found) relative to the other detectors. Averages are of all subjects in the *WSC₁₉₅₀* dataset. ‘All stages’ is the average number of spindles found per PSG recording all night.

	Sleep Stage						
	W	N1	N2	N3	REM	Unscored	All stages
A2	11.6 $\pm$ 30.2 (7.7%)	4.5 $\pm$ 8.3 (3%)	123.9 $\pm$ 94.9 (82.2%)	7.1 $\pm$ 16.9 (4.7%)	1.0 $\pm$ 2.3 (0.7%)	2.6 $\pm$ 6.8 (1.7%)	150.7 $\pm$ 159.4
A3	251.5 $\pm$ 393.3 (13.9%)	86.0 $\pm$ 85.5 (4.7%)	1294.4 $\pm$ 412.0 (71.3%)	97.8 $\pm$ 138.8 (5.4%)	32.1 $\pm$ 41.7 (1.8%)	52.4 $\pm$ 97.7 (2.9%)	1814.2 $\pm$ 1169.0
A4	129.0 $\pm$ 238.0 (13.6%)	43.6 $\pm$ 50.7 (4.6%)	686.2 $\pm$ 183.3 (72.4%)	47.7 $\pm$ 69.6 (5.0%)	14.5 $\pm$ 23.1 (1.5%)	27.3 $\pm$ 58.6 (2.9%)	948.3 $\pm$ 623.3
A5	97.1 $\pm$ 185.0 (11.9%)	35.8 $\pm$ 42.5 (4.4%)	611.4 $\pm$ 188.9 (75.1%)	37.9 $\pm$ 56.5 (4.7%)	11.0 $\pm$ 17.1 (1.4%)	20.5 $\pm$ 46.7 (2.5%)	813.7 $\pm$ 536.7
A7	98.6 $\pm$ 128.2 (11.5%)	46.9 $\pm$ 52.0 (5.5%)	639.8 $\pm$ 480.6 (74.5%)	26.1 $\pm$ 48.8 (3.0%)	25.3 $\pm$ 34.0 (2.9%)	22.2 $\pm$ 30.7 (2.6%)	858.9 $\pm$ 774.3
A7*	17.8 $\pm$ 25.7 (2.6%)	27.5 $\pm$ 30.7 (4.0%)	591.5 $\pm$ 431.1 (86.9%)	26.0 $\pm$ 48.7 (3.8%)	15.1 $\pm$ 21.8 (2.2%)	3.0 $\pm$ 7.9 (0.4%)	680.9 $\pm$ 573.8